

Research Report

ETS RR-14-39

Estimating Item Difficulty With Comparative Judgments

Yigal Attali

Luis Saldivia

Carol Jackson

Fred Schuppan

Wilbur Wanamaker

December 2014

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Research Scientist

Heather Buzick
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Managing Research Scientist

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Donald Powers
Managing Principal Research Scientist

Gautam Puhan
Senior Psychometrician

John Sabatini
Managing Principal Research Scientist

Matthias von Davier
Senior Research Director

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Stellhorn
Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

Estimating Item Difficulty With Comparative Judgments

Yigal Attali, Luis Saldivia, Carol Jackson, Fred Schuppan, & Wilbur Wanamaker

Educational Testing Service, Princeton, NJ

Previous investigations of the ability of content experts and test developers to estimate item difficulty have, for the most part, produced disappointing results. These investigations were based on a noncomparative method of independently rating the difficulty of items. In this article, we argue that, by eliciting comparative judgments of difficulty, judges can more accurately estimate item difficulties. In this study, judges from different backgrounds rank ordered the difficulty of SAT[®] mathematics items in sets of 7 items. Results showed that judges are reasonably successful in rank ordering several items in terms of difficulty, with little variability across judges and content areas. Simulations of a possible implementation of comparative judgments for difficulty estimation show that it is possible to achieve high correlations between true and estimated difficulties with relatively few comparisons. Implications of these results for the test development process are discussed.

Keywords Test development; item difficulty; raters

doi:10.1002/ets2.12042

In Embretson's (1983) conceptualization of construct validity, construct representation concerns identifying the theoretical mechanisms (i.e., the processes, strategies, and knowledge) that underlie test performance and, thus, support the interpretation of test scores. Such representations establish the basis for interpretation of test scores but require a scientific and theoretical foundation for item and test design principles (Embretson, 1998). That is, scientific evidence and theory are needed on how test takers use their knowledge, skills, and abilities when they interact with test items. A better understanding of the relevant processes guiding the test taker in arriving at a response to an item may increase the accuracy of prediction of an item's psychometric features. Principles for test design are emerging for some item types, including popular test items such as paragraph comprehension (Freedle & Kostin, 1993; Gorin & Embretson, 2006) and mathematical problem solving (Enright, Morley, & Sheehan, 1999; Singley & Bennett, 2002), as well as other item types on ability tests, such as spatial items (Bejar, 1993) and nonverbal reasoning items (Embretson, 1998). Nevertheless, a much larger foundation is needed to support test meaning from this kind of evidence (Embretson, 2007).

One source of evidence that could potentially support these investigations is experts' intuitive judgments of item difficulty. These judgments are an indispensable part of test development, as they constitute an important aspect of the appropriateness of an item within the set of other available items and the conceptual framework for the test. Test developers are often explicitly targeting a specific range of difficulty while developing an item because different sets of knowledge, skills, and abilities are measured in different ranges. Surprisingly, investigations of the ability of experts to estimate item difficulty have generally not found much success. Thorndike (1982) asked judges to estimate absolute difficulty on a 9-point scale (with extreme points being *would be passed by no more than 30% of examinees* and *would be passed by 75% or more of examinees*) on verbal analogies, quantitative relations, and figure analogies items. He estimated correlations of .83, .74, and .72 between empirical item difficulty (p^+) and average ratings of 20 raters. Using an application of the Spearman-Brown prediction formula,¹ these correlations for average ratings translate to single judge correlations of .23 to .32. Bejar (1983) provided experts with supporting materials in the form of typical distributions of difficulty for every item type and asked them to rate item difficulty. He found correlations of .15 to .49 with p^+ across raters and item types. In the context of item-level standard setting, several studies report correlations between ratings of item difficulty and actual item difficulty. Melican, Mills, and Plake (1989) found correlations of .26 and .27 between ratings of difficulty of mathematics items and p^+ . Cross, Impara, Frary, and Jaeger (1984) also reported low correlations in the range of 0 to the low 30s.

In another study, Hambleton, Sireci, Swaminathan, Xing, and Rizavi (2003) addressed what they view as a major shortcoming of previous research—the lack of a frame of reference for judgments. They developed and field-tested two methods

Corresponding author: Y. Attali, E-mail: yattali@ets.org

that rely on auxiliary items with known p^+ as an aid in the estimation of the difficulty of other items. In their anchor-based method, judges first discussed attributes of easy, medium, and hard items (defined in terms of three threshold points on the p^+ scale: 25%, 50%, and 75%) and were then provided with two representative items in each of the three difficulty bands. In one study of reading comprehension items, these anchor items were used to help rate the difficulty of 21 other items. The average correlation between the ratings and empirical p^+ was .32 (Hambleton et al., 2003, Table 2). The judges rated the difficulties of the items again after a group discussion of their initial ratings. The average correlation between the revised ratings and p^+ was .44. In a second study with 18 analytical reasoning items, the average correlation between the ratings and p^+ was .37 and .50 for the initial and revised ratings (Hambleton et al., 2003, Table 6). In their item-mapping method, an entire test was presented with p^+ values as an aid for rating new items. With this method, six judges rated 21 logical reasoning items, and the average correlation between the ratings and p^+ was .61 and .76 for the initial and revised ratings (Hambleton et al., 2003, Table 4).

The item-mapping method seems to have produced significantly better results than the anchor-based method and conventional rating methods. This method also has more in common with comparative judgment methods than with independent rating, because it essentially requires placing the item in an appropriate position within a presorted list, instead of independently rating its difficulty. In one early study of item difficulty judgments, Lorge and Kruglov (1953) asked experts to rank order the difficulty of 45 arithmetic items. Results of this study are comparable to the results of the item-mapping method, with rank order correlations that varied between the .50s and .80s, with an average of around .70.

In this article, we argue that the main reason for the inability of experts to judge the difficulty of test items accurately is the use of noncomparative methods for eliciting judgments of difficulty. With the rating task, experts are asked to provide a direct measurement of the difficulty of an item, independently of other items. Therefore, this task assumes that the judge possesses a scale that can be used to perform this measurement. However, as Hambleton et al. (2003) argued, even experts lack a frame of reference for this kind of judgment. This lack of a mental scale means that the difficulty of items can be best judged in comparison to the difficulty of other items. Measurement through methods of comparative judgment has a long history in psychology (Thurstone, 1927). The method of paired comparisons is the most flexible as a basis for scaling, but it is less efficient than rank ordering a larger set of items. In this article, we set forth to explore a comparative judgment method for item difficulty that is based on rank ordering. Rankings of item difficulty by test developers on a large pool of SAT[®] mathematics items were collected and analyzed. Then, a simulation was conducted to estimate the success of one possible method for generating difficulty judgments that are based on item comparisons—asking judges to compare the difficulty of a new item to a series of anchor items with known difficulty.

Method

To explore the comparative judgment method for item difficulty, we asked a group of mathematics test developers to rank order sets of items by difficulty. After some initial experimentation, we decided that rank ordering seven items provides a good balance between cognitive load and efficiency. A full ordering of seven items indirectly provides information about 21 paired comparisons (the first item is compared with six other items, the second with five others, etc.) and can be accomplished relatively quickly.

Materials

Eight major content areas from the SAT mathematics section were selected for analysis: (a) numbers and operations with integers, (b) numbers and operations with real numbers, (c) algebraic translations, (d) algebraic problem solving, (e) algebraic functions, (f) geometry—triangles, (g) coordinate geometry, and (h) data analysis. In each content area, a sample of 28 released multiple-choice items was selected. For each item considered for this study, a measure of the item difficulty, the equated delta,² was available. During item selection, the easiest and hardest items were oversampled to increase the likelihood of all types of comparisons across the difficulty spectrum (e.g., to ensure that judges would compare easy items with easy, medium, and hard items). As a consequence, the delta distribution of the 224 selected items had a negative kurtosis (−1.0), but it was symmetric (skewness of 0.0), with a mean of 12.0 and a standard deviation of 3.6. The items in each content area were arranged in booklets in random order (from 1 to 28), such that each successive set of seven items occupied two pages.

Participants

A total of 26 Educational Testing Service (ETS) test developers and external (to ETS) item writers participated in the study. Participants had different backgrounds and levels of experience: six SAT test developers, five GRE® test developers, 10 experienced item writers, and five relatively new item writers. The SAT test developers were both familiar with SAT items and are regularly exposed to item difficulty indices. GRE test developers are regularly exposed to item difficulty indices but lack experience with SAT item types and the SAT population of examinees. Item writers are not regularly exposed to item indices, but experienced item writers are familiar with the SAT item types.

Procedures

Raters had to write down the item numbers in each set in the order of judged difficulty, from easiest to hardest. In seven cases, raters accidentally repeated one of the item numbers and skipped another item. It was decided not to analyze the results of these sets.

The instructions for test developers were as follows:

For this study, we have assembled eight packets of items in different content areas. Each packet contains 28 items. For each packet, we ask that you consider the following four sets of items separately: 1–7, 8–14, 15–21, and 22–28. For each set, your task is to rank the seven items in estimated order of difficulty. Typically, you would solve each item, form an impression of its difficulty, and compare it to previous items. Sometimes it is easier to form partial rank orders (for sets of similar items) and finally combine the seven items into a single order.

Analyses

Two types of analyses were conducted. First, as an initial descriptive analysis of the rankings, rank-order correlations between difficulty judgments and actual item difficulties (equated delta values) were computed for each set of items. However, these correlations cannot be used as unbiased estimates of the accuracy of comparative judgments because the items in the sets were not randomly selected from a large bank of items. To better estimate the accuracy of rater judgments, as well as the variability across content areas and raters, an analysis of individual paired comparisons was conducted. The complete ranking of each set of seven items produced 21 paired comparisons (7 times 6 divided by 2). In each of these comparisons, the rater either correctly identified the harder item or not. The primary predictor of this binary outcome is the difference in empirical difficulty values (delta) of the two items (in absolute value).

A two-level cross-classified hierarchical general linear model (Raudenbush & Bryk, 2002) was estimated for these data. At level 1, the outcome Y_{ijk} for an individual comparison i of rater j for content area k is assumed to have a Bernoulli distribution with probability of success φ_{ijk} . Traditionally, it is the log of the odds of success that is modeled:

$$\eta_{ijk} = \log \left(\frac{\varphi_{ijk}}{1 - \varphi_{ijk}} \right).$$

If the probability of success is .5, the odds of success are $.5/.5 = 1.0$ and the log-odds, or *logit*, is $\log(1) = 0$.

The linear structural model at level 1 is simply

$$\eta_{ijk} = \beta_{1jk} D_{ijk},$$

where D_{ijk} is the empirical difference in difficulty for the individual comparison, and β_{1jk} is the regression coefficient relating difference in difficulty and probability of success. Note that the structural model does not have an intercept coefficient, because it is assumed that a comparison between two items with the same difficulty will result in a probability of *success* of .5 and a logit of 0.

The level-2 model represents the variation across two random factors, raters and content areas. These factors are conceived as random because it is assumed that the specific raters and content areas in this study are just a random sample from a much larger universe of possible raters or content areas. Variation of slope coefficients is attributable to rater effects and content effects and, in addition, possible rater and content predictors. Only one possible predictor was examined. Based on the background of raters, three dummy variables were created, GRE (whether the rater was a GRE test

developer or not), EIW (experienced item writer or not), and NIW (new item writer or not). The initial level-2 model was

$$\begin{aligned}\beta_{1jk} &= \theta_1 + b_{10j} + c_{10k} + \gamma_{11} (GRE)_j + \gamma_{12} (EIW)_j + \gamma_{13} (NIW)_j \\ b_{10j} &\sim N(0, \tau_{b10}) \\ c_{10k} &\sim N(0, \tau_{c10}),\end{aligned}$$

where θ_1 is the expected slope when all dummy variables are 0 (that is, for an SAT rater), b_{10j} is the random effect of rater j , and c_{10k} is the random effect of content k .

Results

Rank Correlations Between Delta Values and Rater Rank Ordering

Table 1 presents descriptive statistics about the Spearman rank-order correlations between difficulty judgments and equated delta values. Correlations were computed for each set of seven items, with a total of 32 sets (four sets in each of eight content areas) per rater. The expected number of sets is 104 (four sets and 26 raters), but due to the missing values discussed above, some content areas are missing one or two sets. The overall median correlation was .79, with a somewhat lower average correlation of .70, due to a small number of very low correlations. Table 1 also shows some variability in results across content areas (the standard deviation of the median correlation across the eight content areas was .06), with lower results for translations and triangles. Interestingly, test developers predicted that functions and data analysis would be more difficult for raters. Some variability of results also occurred across raters (not shown in Table 1); the standard deviation of the median correlations across the 26 raters was .06.

Paired Comparison Analysis

In this section, we present the results for modeling the probability that a rater correctly judges the harder of two items, as a function of the difference in the actual difficulty between the two items. In the initial two-level cross-classified hierarchical general linear model described in the “Method” section, none of the parameters for the three dummy variables for rater background were statistically significant, suggesting that rater background did not have an effect on the quality of discrimination between items. Therefore, a revised model without rater background is shown in Table 2, which presents the estimates of level-2 coefficients. It shows that the average (or intercept) slope for D is .370, with an odds ratio of nearly 1.5, which means that, for the average rater and content, the odds of a successful comparison increases by half for each increase of 1 in the delta difference between the two items.³ The standard deviation of the slope across raters was .061, and the standard deviation of the slope across content was .055.

Figure 1 presents (the solid line) the expected probability of success for an average rater, as a function of the delta difference between the items. The dotted lines represent the success probabilities for a particularly low (1 standard deviation below the average) and high (1 standard deviation above the average) discriminating rater. Predicted probabilities of success were converted from predicted log-odds ($\hat{\eta}_{jk}$) by computing

$$\hat{\phi}_{jk} = \left(\frac{1}{1 + \exp\{-\hat{\eta}_{jk}\}} \right).$$

For example, if the delta difference is 2, then the predicted logit ($\hat{\eta}$) for an average rater and content is 2 times .370, or .740, and the predicted probability of success is $1/(1 + \exp\{-.740\}) = .677$.

Simulation of Implementation

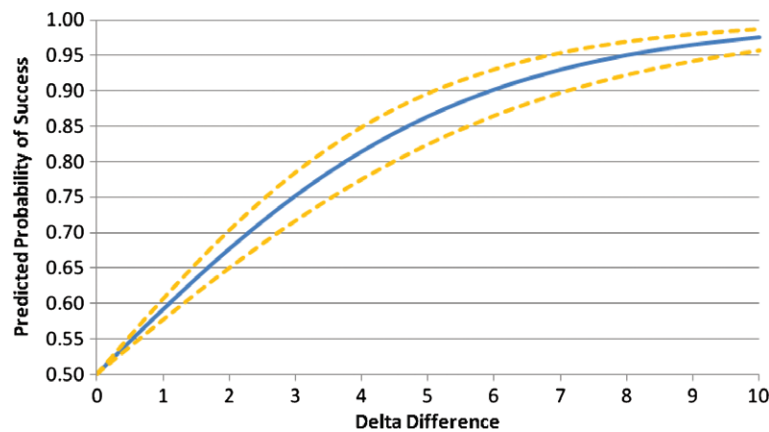
Results of the empirical study showed that raters are reasonably successful in rank ordering a few items in terms of difficulty. A possible method for generating difficulty judgments that are based on item comparisons is to ask raters to compare the difficulty of a new item to a series of anchor items with known difficulty. An efficient set of comparisons (similar to

Table 1 Summary Statistics of Rank Order Correlations Between Empirical Difficulty and Rater Judgments

Content area	<i>N</i>	Median	Mean	<i>SD</i>	5th Percentile
Integers	102	.82	.75	.25	.21
Real numbers	102	.85	.78	.20	.39
Translations	102	.71	.62	.27	.04
Problem solving	104	.77	.71	.21	.32
Functions	104	.77	.68	.29	.21
Triangles	103	.68	.59	.27	.07
Coordinate geometry	104	.82	.78	.15	.43
Data analysis	104	.75	.72	.21	.25
All	825	.79	.70	.24	.21

Table 2 Estimates of Level 2 Coefficients

Fixed effect	Coefficient	<i>SE</i>	Approx. <i>df</i>	<i>t</i> -ratio	Odds ratio
For <i>D</i> slope, β_1					
Intercept, θ_1	.370	.024	17,239	15.7	1.45
Random effect	<i>SD</i>	<i>df</i>	χ^2	<i>p</i> -Value	
Raters, b_{00j}	.061	25	156.3	<.001	
Content, c_{00k}	.055	7	117.5	<.001	

**Figure 1** Expected probability of success for an average rater. Dotted lines represent probability of success for raters 1 SD above and 1 SD below average.

a binary search algorithm) would start with an anchor item with median difficulty, an item at the 50th percentile of the distribution of item difficulty. If the new item is judged more (less) difficult than the initial anchor item, a second anchor item at the 75th (25th) percentile of the distribution can be presented for comparison. Depending on the first two comparisons, a third anchor item at the 12.5, 37.5, 62.5, or 87.5 percentile could be presented. In this manner, paired comparisons can be translated into judgments of difficulty. Each new comparison provides an opportunity to associate the item with a higher or lower level of difficulty. The total number of levels of judged difficulty would be equal to 2^k , where k is the number of comparisons. With three comparisons, the final estimates of the difficulty of the item could be at the 6.25, 18.75, 31.25, 43.75, 56.25, 68.75, 81.25, or 93.75 percentile of the distribution. In addition, more than one rater can be asked to perform this process, so that judgments of different raters could be averaged. Note that, in practice, this method can be implemented as either a series of binary judgments (as described above) or as a simultaneous judgment process, whereby all anchor items are presented at once and the judge needs to select the region for the new item.

To evaluate this method, a simulation was performed. In this simulation, 10,000 items were systematically drawn from a (normal) distribution of item difficulties, and each item was compared to a series of anchor items in the manner described

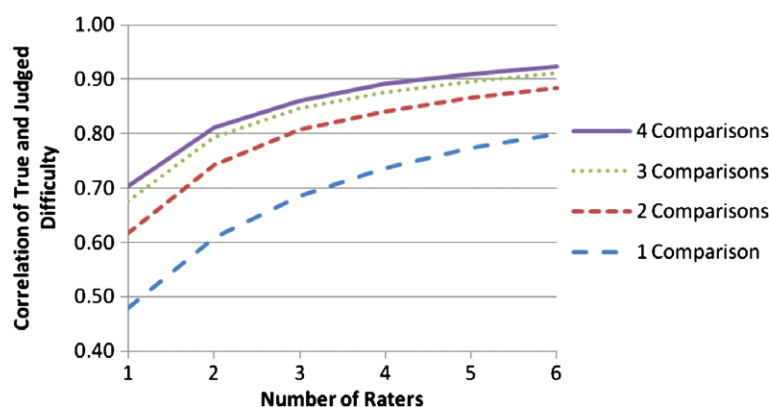


Figure 2 Correlations between true and judged difficulty by number of raters.

above. The result of each comparison was randomly determined based on the probability of success estimated in the previous section. To estimate this probability, the difference in difficulty between the two items (new item and anchor item) was computed and an *average* rater was assumed to compare the two items (one with a slope of .370). Thereafter, a random number from a uniform distribution in the range 0–1 was generated and was compared to the estimated probability of success. If the random number was smaller than this probability, the result of the simulated comparison was successful.⁴

The simulations were repeated with one to four comparisons and with one to six raters, for a total of 24 simulations. For each simulation, the correlation between true difficulty and average judged difficulty level was computed for the 10,000 simulated items. Figure 2 presents these correlations for each simulation. The figure shows that even with a single comparison and a single rater, the resulting two levels of judged difficulty (more or less difficult than the anchor item) result in a correlation of almost .5 with true item difficulty. The figure shows that the added value of increasing the number of comparisons diminishes beyond three comparisons. With respect to the number of raters, a significant increase in correlations can be seen even beyond four or five raters. As an example, with three comparisons, the use of two raters results in a correlation of .80, and the use of five raters results in a correlation of .90.

These correlations can be compared to expected correlations between true and empirical difficulty estimates, obtained from a sample of examinees in an item tryout. With reasonable assumptions about the range of p^+ values, it can be shown⁵ that, for a sample size of 100 examinees, this correlation will be around .90, and for a sample size of 200 examinees, it will be around .95. In other words, the simulation above showed that five raters making comparisons with three anchor items could replicate the accuracy of p^+ estimates with 100 examinees.

Discussion

In this article, we showed that, contrary to previous investigations, judges are able to discriminate quite well between easier and harder items when they are given a comparative judgment task. An interesting result of this study is the relatively small variability across content areas and raters, in the quality of judgments that were generated. In fact, the general linear model results showed no statistically significant differences between the different groups of raters. That is, SAT raters who are most familiar with the items and are regularly exposed to item statistics did not perform better than, for example, new item writers who are not familiar with the items and are not exposed to item statistics. This is an important result, as it suggests that the ability to discriminate between the difficulties of items is less related to test development experience and to experience with a particular difficulty scale.

In an effort to test the limits of this premise, the 14-year-old son of one of the authors was asked to perform the difficulty ranking task. Although familiar with multiple-choice questions, he had never solved SAT questions before. Nevertheless, the rank order correlations for his judgments were .60.

A possible method for generating difficulty judgments that are based on item comparisons is to ask raters to compare the difficulty of a new item to several anchor items with known difficulties. The simulations reported above have used a sequential binary judgment task, because it was easier to model the results of binary decisions. However, it is possible to implement this approach as a single mapping task, similar to the one used by Hambleton et al. (2003). For example, a series of two comparisons requires three items to implement, one middle difficulty item, a second easier item, and a third

harder item. Three comparisons will require seven items, and four comparisons will require 15 anchor items. The above results suggest that three comparisons or seven anchor items may be enough to produce accurate judgments of difficulty, especially if the judgments of more than one rater are averaged. Naturally, the choice of anchor items is important. Some items are more difficult to judge than others, and these should not be used as anchor items.

The ability to predict item difficulty, either judgmentally or through an analysis of item features, also can have practical applications in the test-development process. One possible advantage of obtaining estimates of item difficulty in the process of test development is lowering the sample sizes required for item pretesting, leading to lower costs and increased security of items. Mislevy, Sheehan, and Wingersky (1993) and Swaminathan, Hambleton, Sireci, Xing, and Rizavi (2003) showed how estimates of item statistics could be improved by combining empirical item statistics from a reduced sample of test takers with information on item parameters available from other sources, such as judgments of content experts or theories about the skills and knowledge needed to solve different items. However, these suggestions remained largely hypothetical, given the difficulty of obtaining such additional information. This study shows that, with a comparative task, the prospect of gaining substantial benefits from judgmental predictions of item difficulty is possible. An additional benefit of comparative judgments in this context is that, with comparative judgments, the difficulty estimations are not likely to show systematic biases in judgments. On the other hand, with noncomparative ratings of difficulty, the threat of systematic biases is constant; it cannot be overcome by averaging judgments from several raters, and considerable training is needed to overcome it.

This study focused on mathematics problem-solving items. An interesting issue for future research is the generalizability of these results to other content areas, especially different types of verbal reasoning items. A special complication arises with items that naturally appear in sets, such as reading comprehension items. For these items, it could be more difficult to compare items from different sets, and, therefore, it could be more difficult to find anchor items with existing item difficulty estimates.

Notes

- 1 The correlation between average ratings and p^+ can be assumed to be an estimate of the correlation between true and observed scores. It follows that the reliability of average ratings (e.g., .83 for 20 ratings) is the squared value of this correlation (.69 for the above case). The predicted reliability of an individual rating can then be estimated using the Spearman-Brown formula (.10 for the above case), and the predicted correlation between a single rating and p^+ is the square root of this predicted reliability (.32 for the above case).
- 2 The delta metric is obtained from the p^+ through the inverse normal transformation and is scaled to have a mean of 13 and a standard deviation of 4. On this scale, delta increases for more difficult items.
- 3 The adjusted R -square (Nagelkerke, 1991) of the regular logistic regression model was .49, suggesting D explains approximately half of the variance in rater decisions.
- 4 Note that this simulation assumes a constant discrimination power along the difficulty continuum. In separate logistic regression analyses for easy, medium, and hard items, we did not find evidence to the contrary.
- 5 The correlation between empirical (x) and true (t) p^+ values is given by

$$r_{xt} = \sqrt{\sigma_t^2 / \sigma_x^2} = \sigma_t^2 / (\sigma_t^2 + \sigma_e^2),$$
 and the error variance is equal to the average, across items, of

$$\sigma_e^2 = p(1-p)/N,$$
 where p is the true p^+ for the item, and N is the sample size for item tryout. For a collection of items with p^+ ranging from .2 to .8, we can assume that σ_t is about .15 and that the average for $p(1-p)$ values is around .24 (corresponding to values for a p^+ of .4 or .6).

References

- Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7, 303–310.
- Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–359). Hillsdale, NJ: Erlbaum.
- Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement*, 21, 113–129.

- Embretson, S. E. (Whitely). (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179–197.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 380–396.
- Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, 36, 449–455.
- Enright, M. K., Morley, M., & Sheehan, K. M. (1999). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education*, 15, 49–74.
- Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing*, 10, 133–170.
- Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, 30, 394–411.
- Hambleton, R. K., Sireci, S. G., Swaminathan, H., Xing, D., & Rizavi, S. (2003). *Anchor-based methods for judgmentally estimating item difficulty parameters* (LSAC Research Report 98-05). Newtown, PA: Law School Admission Council.
- Lorge, I., & Kruglov, L. (1953). The improvement of estimates of test difficulty. *Educational and Psychological Measurement*, 13, 34–46.
- Melican, G. J., Mills, C. N., & Plake, B. S. (1989). Accuracy of item performance predictions based on the Nedelsky standard setting method. *Educational and Psychological Measurement*, 49, 467–478.
- Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, 30, 55–78.
- Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691–692.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.
- Singley, M., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 361–384). Mahwah, NJ: Lawrence Erlbaum.
- Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, 27, 27–51.
- Thorndike, R. L. (1982). Item and score conversion by pooled judgment. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 309–326). New York, NY: Academic.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273–286.

Suggested citation:

Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). *Estimating item difficulty with comparative judgments* (ETS Research Report No. RR-14-39). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12042

Action Editor: James Carlson

Reviewers: Kathleen Sheehan and Lixiong Gu

ETS, the ETS logo, and LISTENING. LEARNING. LEADING. are registered trademarks of Educational Testing Service (ETS). SAT is a registered trademark of the College Board. All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at <http://search.ets.org/researcher/>